Artificial Intelligence and Online Extremism: Challenges and Opportunities

Book Section

Artificial Intelligence and Online Extremism: Challenges and Opportunities

Radicalisation is a process that historically used to be triggered mainly through social interactions in places of worship, religious schools, prisons, meeting venues, etc. Today, this process is often initiated on the Internet, where radicalisation content is easily shared, and potential candidates are reached more easily, rapidly, and at an unprecedented scale (Edwards and Gribbon, 2013; Von Behr et al., 2013).

In recent years, some terrorist organisations succeeded in leveraging the power of social media to recruit individuals to their cause and ideology (Farwell, 2014). It is often the case that such recruitment attempts are initiated on open social media platforms (e.g., Twitter, Facebook, Tumblr, YouTube) but then move onto private messages and/or encrypted platforms (e.g., WhatsApp, Telegram). Such encrypted communication channels have also been used by terrorist cells and networks to plan their operations (Gartenstein-Ross and Barr).

To counteract the activities of such organisations, and to halt the spread of radicalisation content, some governments, social media platforms, and counter-extremism agencies are investing in the creation of advanced information technologies to identify and counter extremism through the development of Artificial Intelligent (AI) solutions (Correa and Sureka, 2013; Agarwal and Sureka 2015a; Scrivens and Davies, 2018).

These solutions have three main objectives: (i) **understanding** the phenomena behind online extremism (the communication flow, the use of propaganda, the different stages of the radicalisation process, the variety of radicalisation channels, etc.), (ii) automatically **detecting** radical *users* and *content*, and (iii) **predicting** the adoption and spreading of extremist ideas.

Despite current advancements in the area, multiple challenges still exist, including: (i) the **lack of a common definition** of prohibited radical and extremist internet activity, (ii) the **lack of solid verification of the datasets** collected to develop detection and prediction models, (iii) the **lack of cooperation across research fields**, since most of the developed technological solutions are neither based on, nor do they take advantage of, existing social theories and studies of radicalisation, (iv) the **constant evolution of behaviours** associated with online extremism in order to avoid being detected by the developed algorithms (changes in terminology, creation of new accounts, etc.) and, (v) the development of **ethical guidelines and legislation** to regulate the design and development of AI technology to counter radicalisation.

In this book chapter we provide an overview of the current technological advancements towards addressing the problem of online extremism (with a particular focus on **Jihadism**). We identify

some of the limitations of current technologies, and highlight some of the potential opportunities. Our aim is to reflect on the current state of the art and to stimulate discussions on the future design and development of AI technology to target the problem of online extremism.

An overview of existing approaches

A wide range of work has emerged in the last few years that applied and developed AI technologies with the aim of examining the radicalisation phenomenon, and understanding the social media presence and actions of extremist organisations (Correa and Sureka, 2013; Agarwal and Sureka 2015a; Scrivens and Davies, 2018).

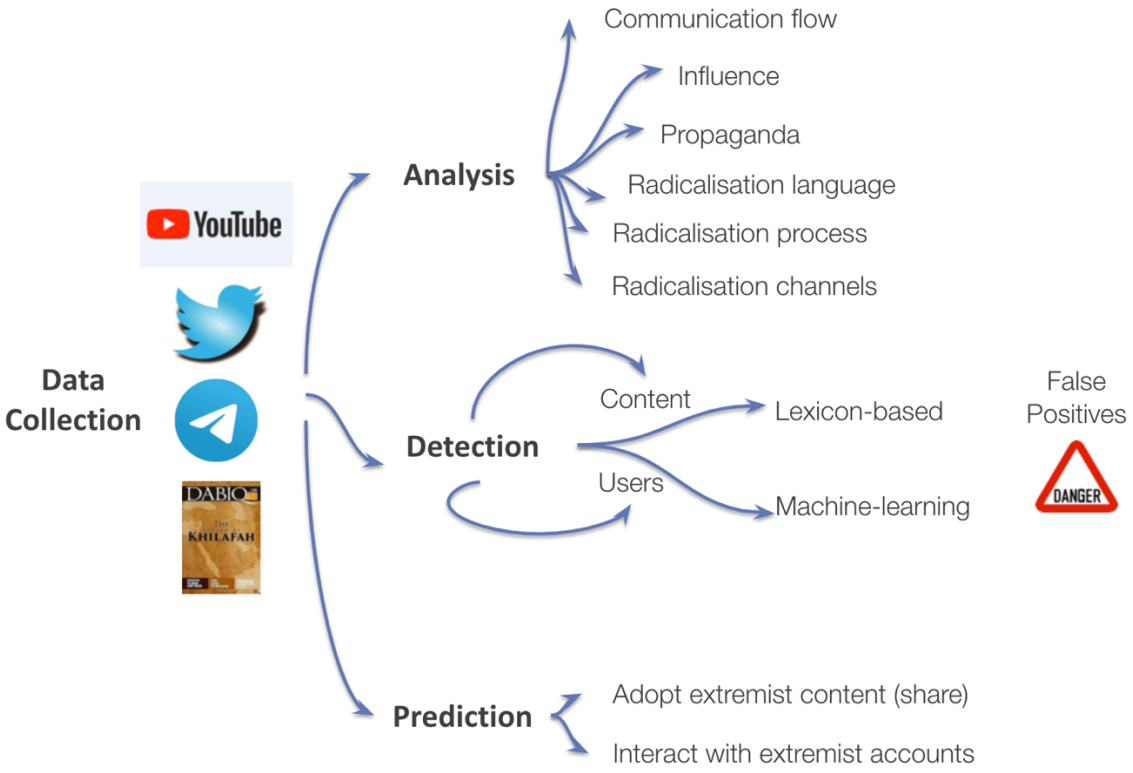

Figure 1: Overview of AI approaches to counter online radicalisation

Broadly, these works can be categorised as (see Figure 1): (i) those that focus on the intelligent, large-scale **analysis** of online radicalisation to better understand this phenomenon, (ii) those that focus on the automatic **detection** of radicalisation, including the detection of radical *content* online, as well as the detection of radical *user* accounts, and (iii) those that focus on the automatic **prediction** of radicalisation (adoption of extremist content, interaction with extremist accounts, etc.). Note that, while we do not present an exhaustive list of works in this chapter, the following sections aim at providing an overview of some representative approaches, including their main

objective, the data they used to support their research, the key algorithms used, and the main output of their work.

Analysis

Works that have focused on the application of AI technologies for the intelligent, large-scale analysis of radicalisation (see Table 1) have different objectives. Among these objectives we can highlight: (i) studying the **communication flow** within the online medium (Klausen, 2015), (ii) analysing **influence** (Carter et al., 2014), (iii) investigating how **propaganda** is presented and spread online (Chatfield et al., 2015; Badawy and Ferrara, 2018), (iv) observe the evolution of **radicalisation language** (Vergani and Bliuc, 2015), (v) study the **radicalisation process** (Bermingham et al., 2009; Rowe and Saif, 2016), and (vi) the analysis of the different online **radicalisation channels**.

Klausen (Klausen, 2015) studied the role of social media, and particularly Twitter, in the jihadists' operational strategy in Syria and Iraq. During 2014, they collected information on 59 Twitter accounts of Western-origin fighters known to be in Syria, and their networks (followers and followees), leading to a total of 29,000 studied accounts. The 59 original accounts were manually identified by the research team. They used known network metrics, such as degree-centrality, number of followers or number of tweets, to identify the most influential users. The authors also conducted a manual analysis of the top recent posts of influential individuals to determine the key topics of conversation (religious instruction, reporting battle and interpersonal communication), as well as the content of pictures and videos. The study highlights the direction of the **communication flow**, from the terrorist accounts, to the fighters based in the insurgent zones, to the followers in the west, and the prominence of female members acting as propagandist.

Carter (Carter et al., 2014), collected during 12-months information from 190 social media accounts of Western and European foreign fighters affiliated with Jabhat al-Nusrah and ISIS. These accounts were manually identified and comprise both, Facebook and Twitter accounts. The paper examined how foreign fighters receive information and who inspires them. The analysis looked at the most popular Facebook pages by "likes", or the most popular Twitter accounts by "follows", as well as the numbers of comments and shares of different posts. The paper also looked at the word clouds of different profiles, revealing terms like (islamic, Allah, fight, Mujahideen, ISIS, etc.) The paper reveals the existence of **spiritual authorities** who foreign fighters go to for inspiration and guidance.

Chatfield (Chatfield et al., 2015) investigated how ISIS members/supporters used Twitter to radicalise and **recruit** other users. For this purpose, they study 3,039 tweets from one account of a known ISIS "information disseminator". Two annotators categorised those posts manually as: propaganda (information), radicalisation (believes in support of intergroup conflict and violence), terrorist recruitment (enticing others to join in fighting the jihad war) and other. Examples of these tweets and their content is provided as a result of this exercise. The analysis also studied the

frequency and times of posting, indicating highly active users, as well as the network of users mentioned in the tweets, which were manually categorised as: international media, regional Arabic media, IS sympathisers and IS fighters.

Vergani (Vergani and Bliuc, 2015) investigated the evolution of the ISIS's **language** by analysing the text contained in the first 11 issues of Dabiq; the official ISIS internet magazine in English. To conduct their analysis they made use of the Linguistic Inquiry and Word Count (LIWC) text analysis program. Their analysis highlighted: (i) the use of expressions related to achievement, affiliation and power, (ii) a focus on emotional language, which is considered to be effective in mobilising individuals, (ii) frequent mentions of death, female, and religion, which are related to the ISIS ideology and the recruitment of women to the cause, and (iv) the use of internet jargon ("btw", "lol", etc.), which may be more effective in establishing a communication with the youngest generations of potential recruits.

While (Klausen, 2015; Carter et al., 2014; Chatfield et al., 2015) studied the social media behaviour of users once radicalised, Rowe and Saif (Rowe and Saif, 2016) studied the social media actions and interactions of Europe-based Twitter users before, during, and after they exhibited pro-ISIS behaviour. Starting from 512 radicalised Twitter accounts, manually identified in the work of O'Callagan (O'Callagan, 2014), they collected their followers, filtered those based in Europe and determined whether those followers were radicalised based on two hypotheses: (i) use of pro-ISIS terminology, a lexicon was generated to test this hypothesis, and (ii) content shared from pro-ISIS accounts. Their filtering process led to the study of 727 pro-ISIS Twitter accounts and their complete timelines. The study concluded that prior to being **activated/radicalised** users go through a period of significant increase in adopting innovations (i.e., communicating with new users and adopting new terms). They also highlight that social homophily has a strong bearing on the diffusion process of pro-ISIS terminology through Twitter.

Bermingham and colleagues (Bermingham et al., 2009) looked at the user profiles and comments of a YouTube video group whose purpose was "the conversion of infidels" with the aim of assessing whether users were being radicalised by the group and how this was reflected in comments and interactions. They collected a total of 135,000 comments posted by 700 members and 13,000 group contributors. They performed term frequency to observe the top-terms used in the group as well as sentiment analysis over a subset of comments filtered by a list of keywords of interest (Islam, Israel, Palestine, etc.). They also used centrality measures to identify influencers. They observed that the group was mostly devoted to religious discussion (not radicalisation) and that female users show more extreme and less tolerant views.

Badawy and Ferrara, (Badawy and Ferrara, 2018), explored the use of social media by ISIS to spread its **propaganda** and to **recruit** militants. To do so, they analysed a dataset of 1.9 million tweets posted by 25K ISIS and ISIS-sympathizers accounts. They distinguish three different types of messages (violence-driven, theological and sectarian content) and they traced a connection between online rhetoric and events happening in the real world.

In 2017, Lara-Cabrera and colleagues (Lara-Cabrera et al., 2017) translated a set of indicators found in social science theories of radicalisation (feelings of frustration, introversion, perception of discrimination, etc.) into a set of computational features (mostly sets of keywords) that they could automatically extract from the data. They assessed the appearance of these indicators in: (i) a set of 17K tweets from pro-ISIS users provided by Kaggle (Kaggle, 2019), a set of 76K tweets from pro-ISIS users provided by Anonymous[1] and a set of 173K tweets randomly selected by opening the Twitter stream. The authors concluded that, while the proposed metrics showed promising results, these metrics were mainly based on keywords. More refined metrics can therefore be proposed to map social science indicators.

Table1: Approaches that focus on the **analysis** of online radicalisation

Work	**Goal**	**Data**	**AI algorithm / technique**	**Conclusions**
(Klausen, 2015)	Study **the communication flow** in the jihadists' operational strategy in Syria and Iraq	59 pro-ISIS **Twitter** accounts (manually assessed) and their networks (29,000 accounts)	**Social network analysis** in combination with manual analysis of accounts, tweets and images	**Communication** flow, from the terrorist accounts, to the fighters based in the insurgent zones, to the followers in the west. Prominence of female members acting as propagandist
(Carter et al., 2014)	Examine how foreign fighters receive information and who inspires them (**influence**)	190 pro-ISIS **Twitter** and **Facebook** accounts (manually assessed)	Manual annotation and assessment of accounts in combination with **social network analysis**	Existence of **spiritual authorities** who foreign fighters look to for inspiration and guidance
(Chatfield et al., 2015)	Investigate how ISIS members/supporters used Twitter to radicalise and **recruit** other users	3,039 tweets from one account of a known ISIS "information disseminator" (**Twitter**)	**Social network analysis** combined with manual analysis of content	Posts about **propaganda**, radicalisation and terrorist recruitment mentioning international media, regional Arabic media, IS sympathisers and IS fighters.
(Vergani and Bliuc, 2015)	Investigated the evolution of the ISIS's **language**	first 11 issues of **Dabiq**, the official ISIS's internet magazine	**Natural Language Processing based on LIWC** (Linguistic Inquiry and Word Count)	Use **expressions related to achievement, affiliation and power**. Emotional language. Mentions of death female and religion and use of internet jargon

[1] https://foreignpolicy.com/2015/11/13/anonymous-hackers-islamic-state-isis-chan-online-war/

(Rowe and Saif, 2016)	Study Europe-based Twitter users before, during, and after they exhibited pro-ISIS behaviour to better understand the **radicalisation process**	727 pro-ISIS **Twitter** accounts. Categorised as pro-ISIS base on the use of radicalised terminology and sharing from radicalised accounts	Modelling and analysis of diffusion over **time-series** data	Prior to being activated/radicalised users go through a **period of significant increase in adopting innovations** (i.e. communicating with new users and adopting new terms). Social homophily has a strong bearing on the diffusion process of pro-ISIS terminology.
(Bermingham et al., 2009)	Explore the use of sentiment and network analysis to determine whether a YouTube group was used as **radicalisation channel**	135,000 comments and 13,700 user profiles. **YouTube** group manually assessed	**Social network analysis and content analysis** (including the automatic extraction of topics and sentiment)	The group was mostly devoted to religious discussion (not radicalisation). Female users show more extreme and less tolerant views
(Badawy and Ferrara, 2018)	Explored the use of social media by ISIS to spread its **propaganda** and **recruit** militants	1.9 million **Twitter** posts by 25K ISIS and ISIS-sympathizers accounts	**Lexicon-based approach** to classify each tweet into violence, theological, sectarian and others, and an **over-time analysis of tweets** and correlation with real-events	**Violence-driven, theological and sectarian conten**t play a crucial role in ISIS messaging. There is a connection between online rhetoric and events happening in the real world
(Lara-Cabrera et al., 2017)	Translate a set of indicators found in social science models into a set of computational features to identify the **characteristics of users at risk of radicalisation**	17K **Twitter** posts from pro-ISIS users provided by Kaggle (Kaggle, 2019). 76K tweets from pro-ISIS users provided by Anonymous. 173K tweets randomly selected	**Five indicators are modelled based on lexicons** (frustration, negative content, perception of discrimination, negative ideas of Western society and positive ideas about Jihadisim) and their density distribution is observed within the data	The proposed **indicators do indeed characterise radicalised users**. Authors define as the next step the use of these indicators as features to create Machine Learning classifiers for the automatic classification of users at risk of radicalisation

Detection

While in the previous section we discuss examples of works which have attempted to analyse the phenomenon of online extremism, with the aim of understanding the different actors involved, and how the process kickstarts and evolves, in this section we focused on those works who have

attempted to provide technological solutions to automatically **detect** the presence of radical ***content*** and ***users*** online (see Table 2).

Works focused on content have attempted to identify radical material (either text, images or videos), while works focused on users have attempted to automatically identify those social media accounts exhibiting radicalisation signs (using radical rhetoric, sharing radical material, etc.). It is important to highlight here that the automatic detection and categorisation of *users* as radical or extremist is a particularly difficult and sensitive problem, since the wrong categorisation of a user as radical (false positive error) may result in an innocent person being subjected to surveillance or policing investigation. In this section we give an overview of some of these works, focusing on their key objectives, the AI methods applied or proposed, the datasets used to conduct the research, and the key obtained outputs.

In 2013, Berger and Strathearn (Berger and Strathearn, 2013) developed an approach to detect **individuals more prone to extremism** (white supremacy in this case) among those with an interest in violent ideologies. Their approach started by collecting the social networks of twelve known extremists on Twitter (3,542 accounts were collected using this process and a maximum of 200 tweets per account was analysed). Using the 3,542 accounts collected using this method, the work measured three dimensions for each user: (i) their influence (number of times their content was retweeted), (ii) exposure (number of times they retweeted other's content) and (iii) interactivity (by looking for keywords in tweets like DM -Direct Message- or email). They concluded that high scores of influence and exposure showed a strong correlation to engagement with the extremist ideology. Manual analysis of the top 200 accounts was used for evaluating the proposed scoring.

In 2015, Berger and Morgan (Berger and Morgan, 2015) aimed at creating a **demographic snapshot of ISIS supporters on Twitter** and outline a methodology for detecting pro-ISIS accounts. Starting from a set of 454 seed accounts (identified by previous research (Berger and Strathearn, 2013) and recursively obtaining followers of those accounts and filtering them based on availability of the account, robot identification, etc., they obtained a final list of 20,000 pro-ISIS accounts to analyse. They estimated that at least 46,000 pro-ISIS accounts were active (as Dec 2014). They created classifiers from a subset of 6,000 accounts that were manually annotated as ISIS supporters or non-supporters. The authors concluded that pro-ISIS supporters could be identified from their profile descriptions: with terms such as succession, linger, Islamic State, Caliphate State or In Iraq all being prominent. When testing this classifier with 1,574 manually annotated accounts they obtained 94% of classification accuracy. However, profile information was only available for around 70% of accounts.

Saif (Saif, 2017) proposed a semantic graph-based approach to **identify pro vs. anti-ISIS social media accounts**. By using this graph, the authors aimed at capturing the relations between terms (e.g., countries *attacking* ISIS vs. countries *attacked* by ISIS) as well as contextual information based on the co-occurrence of terms. Their work hypothesised that, by exploiting the latent semantics of words expressed in social media content, they could identify additional pro-ISIS and

anti-ISIS signals that could complement the ones extracted from previous approaches. The authors developed multiple classifiers and showed that their proposed classifier, trained for semantic features, outperformed those trained from lexical, sentiment, topic and network features. Evaluation was done on a dataset of 1,132 Twitter users (with their timelines). 566 pro-ISIS accounts, obtained from (Rowe and Saif, 2016) and 566 anti-ISIS users, whose stance was determined by the use of anti-ISIS rhetoric.

Fernandez (Fernandez and Alani, 2018) hypothesise that a key reason behind the inaccuracy of radicalisation detection approaches is their reliance on the appearance of terminologies and expressions regardless of their **context**. The authors therefore explore: (i) how pro-ISIS users and non pro-ISIS users (journalists, researchers, religious users, etc.) use the same words and expressions, (ii) if there exist any divergence in how the same words are used, and (iii) if this context divergence can be helpful **to create more accurate radicalisation detection methods**. The work uses 17K tweets from pro-ISIS users and 122K tweets from 'general' Twitter users available via the Kaggle datasets (Kaggle, 2019). This work concludes that the identification of language divergence between these groups can lead to more accurate user and content detection mechanisms.

Stepping aside from the categorisation of users as ‘radical’ or ‘non-radical’, Fernandez (Fernandez et al., 2018) proposed an approach to **measure the influence of online radicalisation that a user is exposed to**. The proposed approach renders the social science theory of ‘roots of radicalisation’ (Schmid, 2013; Borum, 2016) into a computational model that computes the micro (individual, i.e., originating from the user himself), meso (social, i.e., originating from the user’s social network) and macro (global, i.e., originating from events happening in the world) radicalisation influence a user is exposed to based on her social media contributions. The work used 17K tweets from pro-ISIS users and 122K tweets from 'general' Twitter users available via the Kaggle data science community (Kaggle, 2019), and concluded that there is an important need to leverage more strongly the knowledge of theoretical models of radicalisation to design more effective technological solutions for the tracking of online radicalisation.

Agarwal and Sureka (Agarwal and Sureka 2015b) investigated techniques to automatically identify hate and **extremism promoting tweets**. Starting from 2 crawls of Twitter data they used a semi-supervised learning approach based on a list of hashtags (#Terrorism, #Islamophobia, #Extremist) to filter those tweets related to hate and extremism. The training dataset contained 10,486 tweets. They used random sampling to generate the validation dataset (1M tweets). Tweets were in English and manually annotated by four students. They created and validated two different classifiers (KNN and SVM) based on the generated datasets to classify a tweet as hate promoting or unknown. By creating and validating these classifiers, they concluded that the presence of religious, war related terms, offensive words and negative emotions are strong indicators of a tweet to be hate promoting.

Ashcroft (Ashcroft et al., 2015) investigated the automatic detection of **messages released by jihadist groups** on Twitter. They collected tweets from 6729 Jihadist sympathisers. Two

additional datasets, one of 2,000 randomly selected tweets, and one of tweets from accounts manually annotated as anti-ISIS, were collected for validation. Numbers of tweets for the pro and anti-ISIS datasets were not reported, but based on the provided experiments we estimated they should be around 2,000 each. SVM, Naive Bayes and Adaboost classifiers were trained with this data using stylometric, time and sentiment features. Authors concluded that Fridays are a key date to spread radical tweets and that automatic detection is viable but can never replace human analysts. It should be seen as a complementary way to detect radical content.

Table 2: Approaches that focus on the **detection** of online radicalisation

Work	**Goal**	**Data**	**AI Algorithm / Technique**	**Conclusions**
(Berger and Strathearn, 2013)	Identify individuals prone to extremism from the followers of extremist accounts (**user detection**)	3,542 **Twitter** accounts (followers of 12 known pro-ISIS accounts)	Designed a **scoring system** to measure "**influence**" and "**Exposure**" based on interactions such as replies, retweets, or direct messages	**High scores of influence an exposure showed a strong correlation to engagement** with the extremist ideology (manual evaluation)
(Berger and Morgan, 2015)	Create a demographic snapshot of ISIS supporters on Twitter and outline a methodology for detecting pro-ISIS accounts (**user detection**)	20,000 pro-ISIS **Twitter** accounts (7,574 manually annotated to test classification)	A Machine Learning (ML) **classifier** was trained based on 6,000 accounts and tested with 1574. No details are provided on the ML method used.	The authors concluded that pro-ISIS supporters could be identified from their **profiles descriptions**: with terms such as succession, linger, Islamic State, Caliphate State or In Iraq all being prominent
(Saif, 2017)	Create classifiers able to automatically identify pro-ISIS users in social media (**user detection**)	1,132 **Twitter** users (566 pro-ISIS, 556 anti-ISIS). Annotation based on the terminology used and the sharing from known radicalised accounts	SVM **classifiers** are created based on n-grams, sentiment, topic and network features. The authors also proposed classifier based on semantic features (frequent patterns extracted from a **knowledge-graph**).	Classifiers trained on **semantic features** outperform those trained from lexical, sentiment, topic and network features
(Fernandez and Alani, 2018)	Explore the use of semantic context to create more accurate radicalisation detection methods (**user detection**)	17K tweets from pro-ISIS users and 122K tweets from 'general' **Twitter** users available via the Kaggle data science community (Kaggle, 2019)	Semantic extraction of entities, entity types, topics and categories from a **knowledge graph** (to model context) and incorporation of such context as features into SVM, Naive Bayes and Decision Tree **classifiers**.	**Semantic information can help to better understand** the **contextual variances** in which radicalisation terms are used when conveying 'radicalised meaning' vs. when not. Understanding such variances can help to create more accurate

				radicalisation detection methods.
(Fernandez et al., 2018)	Measure the **influence of online radicalisation** that a user is exposed to. Design a computational method based on the social science theory of roots of radicalisation (Schmid, 2013; Borum, 2016) (**user detection**)	17K tweets from pro-ISIS users and 122K tweets from 'general' **Twitter** users available via the Kaggle data science community (Kaggle, 2019)	Use **word vectors** to model the micro (individual), meso (social) and macro (global) radicalisation influence. **Cosine similarity** is used to compare such vectors against a Lexicon of radical terms	There is an important **need to leverage closer the knowledge of theoretical models** of radicalisation to design more effective technological solutions to track online radicalisation.
(Agarwal and Sureka 2015b)	Automatic identification of hate and extremism promoting tweets (**content detection**)	10,486 hate and terrorism-related **Twitter** posts (extracted based on hashtags) + 1M random tweets annotated by students for validation	They tested KNN and LibSVM **classifiers** based on religious, offensive, slang, negative emotions, punctuations and war related terms	Presence of **religious, war related terms, offensive words and negative emotions** are strong indicators of a tweet to be hate promoting
(Ashcroft et al., 2015)	Automatically detect messages released by jihadist groups on Twitter (**content detection**)	2,000 pro-ISIS **Twitter** posts (containing pro-ISIS terminology and extracted from the accounts 6,729 ISIS sympathisers), 2,000 anti-ISIS tweets(extracted from manually assessed anti-ISIS accounts), 2000 random tweets.[2]	Trained **classifiers** (SVM, Naive Bayes and Adaboost) based on stylometric (n-grams, hashtags, word frequency, etc.), time-based and sentiment features	Fridays are a key date to spread radical tweets. **Automatic detection is viable but can never replace human analysts.** It should be seen as a complementary way to detect radical content

Prediction

Regarding **prediction** of radicalisation (see Table 3), we can highlight the works of (Magdy et al., 2016) and (Ferrara et al., 2016).

Magdy (Magdy et al., 2016) proposed an approach to identify Arab Twitter accounts explicitly expressing positions supporting or opposing ISIS. They collected 57,000 Twitter users who authored or shared tweets mentioning ISIS and determined their stance based on the use of the full name of the group vs. an abbreviated form. They then created classifiers to predict future support

[2] Numbers of pro and anti-ISIS tweets are not reported but estimated based on the experiments

of opposition to ISIS based on the users' timelines before naming ISIS. The authors conclude that Pro- and anti-ISIS users can be identified before they voice explicit support or opposition.

Ferrara (Ferrara et al., 2016) proposed a computational framework for detection and prediction of extremism in social media. For this purpose, they used a dataset of over 3M tweets generated by over 25 thousand extremist accounts, who have been manually identified, reported, and suspended by Twitter (Ferrara, 2017), and a dataset of 29M posts from the followers of these users. Random forest and logistic regression were used for classification and prediction based on user metadata and activity features, time features, and features based on network statistics. Two types of predictions were made: (i) whether the follower will adopt extremist content (retweet from a known pro-ISIS account) and (ii) whether the follower will interact (reply) with a known pro-ISIS account. The authors concluded that the ratio of retweets to tweets, the average number of hashtags adopted, the sheer number of tweets and the average number of retweets generated by each user, systematically rank very high in terms of predictive power.

Table 3: Approaches that focus on the **prediction** of online radicalisation

Work	**Goal**	**Data**	**AI Algorithm / Technique**	**Conclusions**
(Magdy et al., 2016)	Proposed an approach to **predict future support or opposition to ISIS**	57,000 **Twitter** users who authored or shared tweets mentioning ISIS. Categorised as pro or anti-ISIS based on the use of the full name of the group vs. an abbreviated form	SVM **classifier** based on bag-of-words features, including individual terms, hashtags, and user mentions	**Pro- and anti-ISIS users can be identified before they voice explicit support or opposition.**
(Ferrara et al., 2016)	Propose a computational framework for detection and **prediction of: adoption of radical content and interaction with pro-ISIS accounts**	Over 3M **Twitter** posts generated by over 25 thousand extremist accounts (manually identified, reported, and suspended by Twitter). 29M posts from the followers of these accounts	Random forest and logistic regression **classifiers** are used for classification and prediction based on user metadata and activity features, time features, and features based on network statistics	The ratio of retweets to tweets, the average number of hashtags adopted, the sheer number of tweets and the average number of retweets generated by each user, systematically rank very **high in terms of predictive power**

Challenges

Despite the previous advancements in the area, multiple challenges still exist when targeting online radicalisation. These challenges include: (i) the ones that are derived from conducting *research*

with Big Data[3] such as (Volume -large amounts of content-, Velocity -new content quickly produced-, Variety- heterogeneity of the data and the information sources where data is produced- and Veracity -quality of the information-), (ii) the ones that are derived from the application of technology into a new field (such as *technology adoption* by users), and (iii) the ones that are *specific to online radicalisation research and the development of AI applications to counter radicalisation*. Although we acknowledge the challenges derived from the use of big data, and the challenges for relevant stakeholders to adopt novel counter radicalisation technology, in this book chapter we aim to focus on the specific challenges of the design and development of AI solutions to counter radicalisation. We have identified six main challenges in this work (see Figure 2). These challenges are described in the following subsections.

Figure 2: Main challenges of the development of AI applications to counter radicalisation

Defining Radicalisation

One of the key challenges of the design and development of AI technology to target radicalisation is the **lack of a common definition of prohibited radical and extremist internet activity**, which can impede optimal enforcement (Housen-Couriel et al., 2019). Online radicalisation is a global phenomenon, but it is perceived differently in different regions of the world, and hence it is complicated to have a single unique and globally accepted definition (Meserole and Byman, 2019).

Currently, many governments around the world are pressurising globally operating Tech companies, such as Google, Facebook, and Twitter, to remove and block radical content and

[3] http://researchhubs.com/post/ai/introduction-to-data-science/big-data-4-v.html

accounts.[4] However, no clear definitions of what constitutes a radical piece of content, or a radical account are provided with these government regulations, which means that Tech companies have to set up their own definitions and to decide which content they block or which content they keep online, with the corresponding ethical implications that this entails (Saltman, 2019). Initiatives such as the *Global internet forum*[5], or *Tech against Terrorism*[6] have emerged in recent years with the idea of formalising definitions and fostering collaborations among Tech companies, civil society, academics, governments and supra-national bodies such as the European Union (EU) and United Nations (UN). However, more dialog and collaboration across these organisations is needed to reach a consistent definition.

Data collection, verification and publication

Another very important challenge when researching online radicalisation is the availability and quality of data used to study this phenomena.

As we have seen in the previous works, multiple datasets have been collected for studying radicalisation. However, many of these datasets are collected based on certain assumptions (e.g., accounts that use radical terminology or share radical material (Rowe and Saif, 2016; Magdy et al., 2016), accounts that follow known radical accounts (Chatfield et al., 2015), accounts that participate/comment in particular YouTube channels known to disseminate radical content (Bermingham et al., 2009)) but in many occasions, neither those assumptions, nor the data collected based on those assumptions, are properly verified. It is therefore unclear how the amount of noise (content that is not reliable or credible) that exists in those datasets is **affecting the quality and validity of the insights gained from that data** (Parekh et al., 2018). In this section we report on the problems derived from the current mechanisms used for data collection, verification and publication.

[4] https://ec.europa.eu/commission/sites/beta-political/files/soteu2018-preventing-terrorist-content-online-regulation-640_en.pdf

[5] https://www.gifct.org/leadership/

[6] https://www.techagainstterrorism.org/

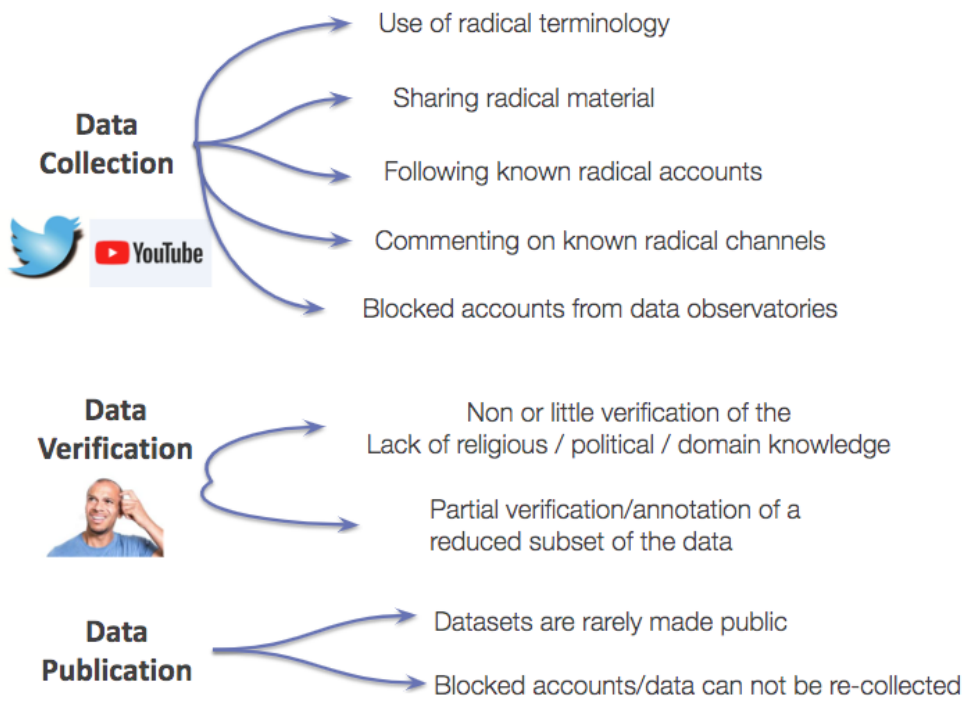

Figure 3: Mechanisms used for data collection, verification and publication

- **Data collection.** As reported in the previous section, common methods used for data collection include: (i) data collected based on the appearance of certain terms ('ISIS', 'daesh', etc.), (ii) data collected based on users sharing a particular URL / image or piece of radical material, (iii) data collected from followers' of known radical accounts, (iv) data collected from users that comment in radical channels (such as YouTube channels), and (v) data collected for accounts that have been blocked or suspended (using data archives gathered by data observatories).
- **Data verification.** Once data is collected based on these assumptions, these data is either not verified, or partially verified, i.e., only a subset of the data is labelled by human annotators. These annotators are generally not experts, but students, or crowdworkers of crowdsourcing platforms (Agarwal and Sureka 2015b). These annotators may not have the religious, political or domain knowledge to assess whether a piece of content, or a particular user account, should be categorised as 'radical'. The other major problem with data verification is the cultural perception. Gold standards have been found to vary depending on who is doing the annotation. In this case the same piece of content may be perceived as radical by experts of certain countries / cultural backgrounds, but may be perceived as non radical in a different cultural / socio-political context. (Patton et al., 2019, Olteanu et al., 2019).
- **Data publication.** Due to the sensibility of the problem, the involvement of personal data, and existing data regulations, such as the General Data Protection Regulation (GDPR) (GDPR, 2019), datasets collected to study radicalisation are not publicly shared. Very few

datasets existing online for research purposes, such as the ones exposed by the Kaggle data science community (Kaggle 2019). It is often the case that researchers do not share the data, and only provide a description of the used data and collection in their papers. However, once content or accounts have been blocked on social media platform, related data cannot be re-collected any longer. It is sometimes possible to retrieve a sample of the blocked content or accounts from data observatories (Ferrara, 2017), but it is unknown what percentage of such information is lost.

In the following subsections we describe some of the challenges derived from the existing data collection, verification and publication mechanisms (see Figure 4).

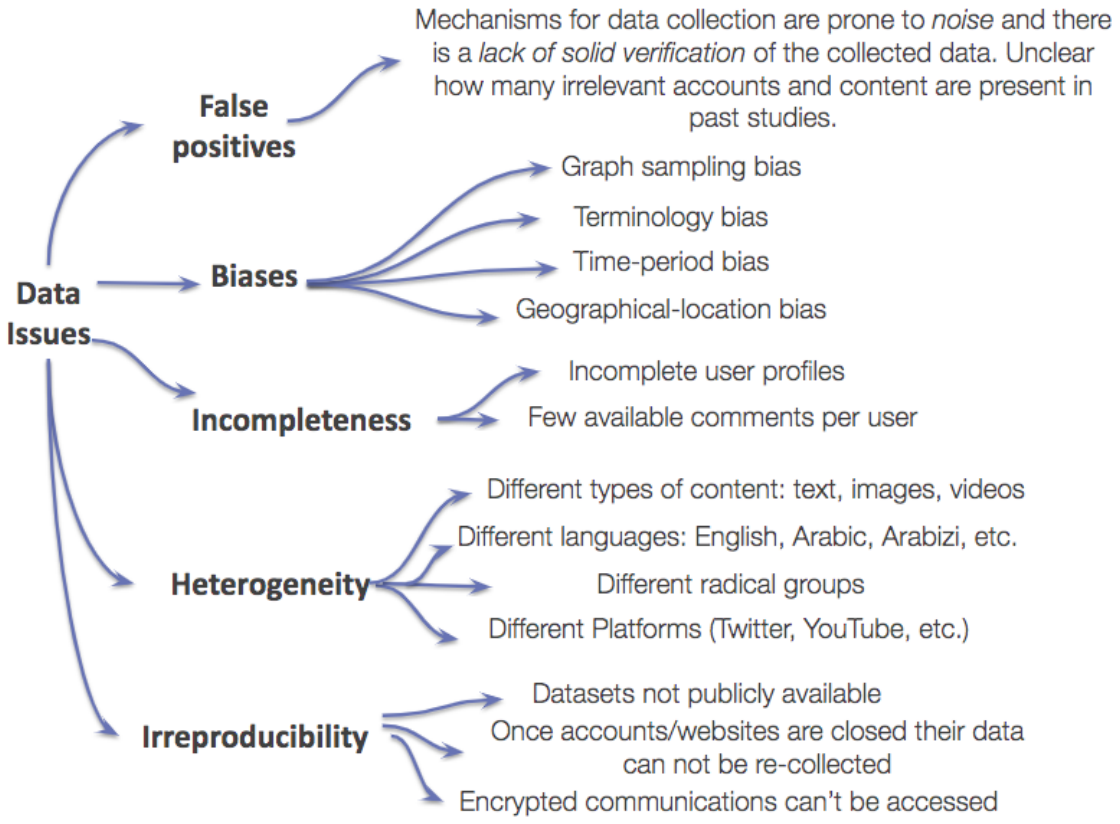

Figure 4: Problems with existing datasets to study radicalisation

Noisy Data (False Positives)

Since existing data collection mechanisms to study online radicalisation are prone to noise, and the collected data is not verified, or only partially verified, the generated datasets could include an unknown amount of false positives (i.e., content and user accounts that, while categorised as radical, are indeed not radical). Examples include content and accounts reporting current events, e.g., "Islamic State hacks Swedish radio station", sharing harmless religious rhetoric, e.g., "If you want to talk to Allah, pray. If you want Allah to talk to you, read the Qur'an", or sharing counter

extremism narratives and material, e.g., "armed Jihad is for defense of Muslim nation, not for stablishment of the Khilafah". Parekh and colleagues (Parekh et al., 2018) highlighted this problem in their analysis and stressed the fact that "*nobody knows how many irrelevant accounts are present in past studies and therefore much of what is known from past studies of online jihadist behavior is highly skewed by irrelevant accounts".*

Learning from noisy data not only means that analysis results may be imperfect, but it also means that the algorithms developed to detect and predict radicalisation may not perform at the reported levels of correctness, since they are trained based on erroneously labelled data. Since irrelevant accounts may erroneously be categorised as 'radical', based on existing data collection methods, it is possible that, while training from these data, the detection and prediction algorithms associate patterns of non-radical accounts (e.g., journalist that report about terrorist-related events, or religious non violent individuals) to the radical category. Note that the erroneous categorisation of a user as radical by a developed AI algorithm may lead to surveillance, or in the worst case, investigation of an innocent individual, which calls for better assessments of different types of errors (Olteanu et al., 2017). Additionally, because the datasets used to train these algorithms are not generally public, it is not possible to verify up to which extent they contain noisy data.

Biases

Existing data collection mechanisms are prone to data biases (i.e., the distortion the sampled data that compromises its representatives) (Olteanu et al., 2019). Collected samples may therefore not be representative of the larger population of interest. Common biases across online radicalisation research include: the terminology bias, the time-period bias, the graph-sampling bias and the geographical location bias. However, it is relevant to note that demographic (age / gender / etc.), as well as behavioural biases (e.g., the various ways in which users connect and interact with one another), may also exist within the collected datasets.

- **Terminology:** When data is collected based on restricted lexicons (i.e., selected terms and expressions), these lexicons may cover only a fraction of the topics or entities (persons, organisations, etc.) discussed by radical groups. They may also cover only the terminology of a particular subgroup, or even only one language (e.g., Arabic). Collected content and accounts are therefore biased to the original lexicons used for collection. It is therefore important to acknowledge that the obtained findings (or developed radicalisation detection classifiers) may not be general but restricted to particular topics of discussion.
- **Time-period:** Data collections are generally restricted to particular time periods (generally a few months). Data is therefore biased to the world events happening during those particular months (i.e., particular terror attacks, regions of the world, political and religious figures, etc.). Classifiers may therefore learn that naming certain political or religious figures, or locations, are reliable indicators to determine whether a piece of content, or a user account, is radical. However, as time evolves, those locations, those popular figures, those events, may not be relevant or even discussed any longer. In certain cases, they may

even become discriminative of the opposite class (e.g., locations under control by a radical group that become liberated) Hence, classifiers trained on data collected in the past to detect and predict radicalisation in the present, or the future, may not perform with the expected level of correctness.

- **Graph-sampling:** The discovery of related accounts is generally based on graph sampling methods, where related accounts are discovered from the social graph of known radical (i.e., seed) accounts. The expanded dataset therefore depends on the choice of the initial seed accounts. The other key problem is the type or relations selected to do graph sampling, since in the case of *'followers'*, irrelevant or noisy accounts, such as reporters and researchers, who are just 'listening' to jihadist accounts, are likely to be included (Parekh et al., 2018; Klausen, 2015).
- **Geographical:** Some data collections are restricted to particular regions of the world (e.g., western countries (Rowe and Saif, 2016)). Variations in user generated content, particularly text, are well-documented across and within demographic groups. Findings about 'radicalisation' based on such samples may hence not be generalisable to other regions in the world.

Incompleteness

When datasets are collected based on keywords, or when data is gathered based on comments on particular YouTube channels, or social media groups, the collected datasets contain very few posts (if more than one) associated to a particular user account (**incomplete user profiles**). This means that, for most user accounts, only a partial view of the history (or timeline) of such account is available. This limits the type of research that can be conducted, since it is very difficult (if not impossible) to study the behavioural evolution of users towards more radical views if their historic posts are not available. Since social media platforms sometimes close radical accounts fairly quickly, recollecting data from such accounts is no longer possible. Similarly, although accounts that get blocked tend to resurface under different names (Conway et al., 2017), those accounts do not have historical data, and therefore AI solutions need to deal with a **'cold start' problem** (i.e., accurate inferences cannot be drawn from accounts for which we have not yet gathered sufficient information).

An additional element of incompleteness within existing datasets are the collected social graphs. In most occasions only a partial sample of the social graph of the collected accounts is being gathered (**incomplete social graphs**). Some researchers tried to reproduce social graphs based on implicit connections (e.g., users mentioning other users within the content (Fernandez et al., 2019)), when the explicit (friend / follower) relations among accounts are not available.

Heterogeneity (Variety of content)

Another relevant consideration is the heterogeneity of the data. Online data comes in multiple languages, from multiple platforms, in multiple formats (audio, video, text) and from multiple

radical groups and subgroups. The development of 'generic' online radicalisation detection methods in an ever changing and heterogeneous world is a complex and challenging task.

- **Language:** Online data comes in multiple languages, sometimes underrepresented languages or forms of text, such as Arabizi (Arabic language written in Latin Script (Tobaili et al., 2019)). Multiple challenges arise when dealing with the multilingual sea of data available online: (i) the lack of resources and local expertise to analyse underrepresented languages, (ii) the automatic identification of the written language, since not only languages coexist across different pieces of content, but the same piece of content may contain terms and expressions in more than one language, and (iii) the informality of social media language, which is an added challenge to the multilinguality of the text. Note that terms and expressions in social media are sometimes written without following standard morphological, or syntactic rules (e.g., 'Heeeeeello' vs. Hello). It is also the case that communities and social groups often invent and adopt terms to define new realities. For example, the acronym KTHHFV, adopted within extreme misogynist communities (Farrell et al., 2019) refers to a Kissless, touchless, hugless, handholdless, friendless, virgin person.[7] Not only those new terms and expressions are not available within standard dictionaries, but also the meanings of those terms and expressions are not known outside the communities that invented and adopted them, hence expert or inside knowledge is needed to capture these complex semantics.
- **Platforms:** radical content is shared in multiple social networking platforms, including Twitter, Reddit, YouTube, Watsapp, Telegram, etc. Each platform differs on how content is posted (e.g, Twitter limits the amount of characters of a posts while other platforms don't have length restrictions) or how user relations are established (e.g., Twitter distinguishes between 'followers - people who follow a user account' and 'followees - people to whom the user account follows') whether others, like Watsapp, do not consider bidirectional relationships. There are also distinctions on how content is shared, how accounts are referred to, or named, whether videos can be streamed, etc.
- **Radical groups:** Not only different accounts may express different extremist ideologies (Jihadist, Far-right, extreme misogyny, etc.), but also within the same extremist ideology we may find different groups. These groups, while having some common ground, differ in their interpretation of concepts, and in their attitudes and actions. Not only these groups coexist within the online world, they also merge and shift depending on real-world events, interests and conflicts.[8]
- **Content types:** In the online world different types of content emerge including videos, images, text, etc. The automatic processing of multimedia content is very different than the processing of textual content. Combinations of AI techniques are therefore needed to understand the complete picture of the radical material being disseminated.

[7] https://www.reddit.com/r/IncelTears/comments/aoekwm/incel_language_dictionary/
[8] https://monitoring.bbc.co.uk/product/c200ntjp

Irreproducibility

A key problem of existing radicalisation research is the lack of reproducibility, since datasets used to study radicalisation are not shared, and once user accounts or content are blocked, data can no longer be recollected.

- **Datasets are not publicly available:** As mentioned when describing existing data publishing methods, while multiple datasets have been collected for research, due to the sensitivity of the data, and to comply with existing social networking sites regulations (Twitter data policies, 2019), and data regulations, such as GDPR (GDPR, 2019), researchers are not sharing the collected datasets. This implies that: (i) further assessments over the data are very difficult to perform, (ii) researchers struggle to build on previous studies and developed systems to further advance research.
- **Once accounts/websites are closed data cannot be recollected:** Researchers working in online radicalisation sometimes share the IDs of forums/groups, accounts or posts (Farrell et al., 2019), so that other researchers can recollect the data. The problem in this case is that, if the collected accounts were indeed radical, or the collected content exhibited radical terminology or material, they will be blocked at the time of recollection. Moreover, according to the data regulations of some social media platforms, like Twitter (Twitter data policies, 2019), researchers and practitioners that collect data are responsible of making all reasonable efforts to delete the collected content, if such content is deleted, gains protected status or is suspended (unless otherwise prohibited by applicable law or regulation, and with the express written permission of Twitter). This regulation makes it even more difficult to maintain datasets to study radicalisation.
- **Encrypted and private communications cannot be accessed:** Extremist organisations sometimes move from the public sphere to a more private medium. This is for example the case of the Islamic State (IS), which moved many of its communications to Telegram due to the disruption they suffered on more visible platforms such as Twitter (Conway et al., 2017). Platforms such as Telegram or Whatsapp offer end-to-end encrypted communications. Therefore, messages sent via private channels, groups and chats cannot be collected. Journalist and researchers have nonetheless gathered and studied information from these platforms via public Telegram channels, and by infiltrating private groups (Clifford and Powell, 2019). IS has also started to experiment with the Decentralised Web. Platforms such as RocketChat and ZeroNet proved attractive for IS media operatives since the developers of those platforms are unable to act against content that is stored on user-operated servers or dispersed across the user community (King, 2019).

Research Methodologies

Various problems and challenges are also derived from the research methodologies used to investigate online radicalisation. We will discuss in this section two common issues: (i) the lack

of a control group to contrast research findings, and (ii) the lack of comparison across existing technological solutions.

Lack of comparison against a control group

One of the key problems with existing radicalisation research is the **lack of comparison against a control group**. Most data analysis approaches are based on the study of datasets containing radical content, or radical accounts (Bermingham et al., 2009; Chatfield et al., 2015; Rowe and Saif, 2016; Badawy and Ferrara, 2018). Based on the analysis of these datasets, these works make conclusions on the most discriminative features or characteristics of radical content and users. However, they do not investigate how these features differ from those of a control group (e.g., religious not violent accounts, accounts from journalist reporting about related events, counter-extremist accounts, and accounts from users with no particular relation with radicalisation). Unless such comparisons are made, it is not possible to claim that certain terms, behaviours, networks, etc., are specific of extremist content or accounts.

In the case of the creation of detection and prediction approaches, most works use a control group, so that the AI algorithms can learn the key discriminative features and divergences between the radical and the non-radical control group. The key problem with some of these works is that, in the majority of the cases, **the used control group is composed by randomly collected posts and user accounts** (Agarwal and Sureka 2015b; Lara-Cabrera et al., 2017). These are the accounts of average social platform users (who may talk about their work, their pets, or other topics not even partially related with extremism or radicalisation). *The key challenge however, lies on differentiating radical accounts from those that, despite using the same terminology, reporting the same events, or talking about the same topics, are indeed not radical* (e.g., accounts or religious not violent individuals, journalist accounts, counter-terrorism accounts, etc.). While some works have attempted to generate control-group datasets by considering similar lexicons to collect radical and non-radical accounts (Fernandez and Alani, 2018), those accounts are not-verified, and it is therefore not possible to determine whether the control group contains representative examples of the above mentioned categories, or simply standard Twitter users that at some point in time, share the same terminology than the radical group under study. Another issue emerges when the control group is collected in a different time period (Lara-Cabrera et al., 2017). Classifiers may then select as discriminative features of the non-radical class terms, like political figures, simply because these terms did not exist in the previous time period when the radical group was collected.

Lack of comparison across approaches

The other main issue that we observe within the literature is the lack of comparison across existing approaches. Different works have analysed and are trained over different datasets, making results and approaches not easily comparable.

In (Correa and Sureka, 2013; Agarwal and Sureka 2015a), the authors conducted an extensive survey of the techniques used to identify and predict radicalisation in social media. From these systematic literature reviews, and the overview provided in this paper, we can observe the use of multiple techniques within different subfields of AI including:

- *Natural Language Processing* (NLP) and the use and development of lexicons to interpret text (Vergani and Bliuc, 2015; Fernandez et al., 2018; Badawy and Ferrara, 2018),
- *Machine Learning* (ML) mostly supervised approaches (SVM, Linear Regressions, Naive Bayes, Decision Trees, and lately deep-learning models) for the automatic detection and prediction of radicalisation (Berger and Morgan, 2015; Agarwal and Sureka 2015b),
- *Semantic Web technologies* (entity and relation extraction and analysis) (Saif et al., 2017; Fernandez and Alani, 2018) to better identify the semantic *context* in which words and expressions are used, or the context in which certain entities (persons, organisations, locations) are mentioned, as a way to improve the accuracy of existing algorithms for radicalisation detection.
- *Information Retrieval techniques* (IR), particularly the use of *ranking* methods and *recommender systems* (Fernandez et al., 2018; Fernandez et al., 2019), as a way to filter and rank content and accounts rather than providing a binary categorisation (radical vs. non-radical)

However, while literature surveys have attempted to identify the wide range of AI techniques used to counter online radicalisation, to the best of our knowledge, **there are no replication studies in the literature attempting to compare existing approaches and techniques**. Comparative studies could help to determine which features, or which classification methods do actually perform more reliably, accurately and efficiently, and under which contexts, when countering online radicalisation. It's also important to note that the algorithms designed and developed by Tech companies (such as Twitter, Google, or Facebook) are not public, and therefore not available for comparison.

Lack of cooperation across research fields

Understanding the mechanisms that govern the process of radicalisation, and online radicalisation in particular, has been the topic of investigation in multiple research fields including: *social sciences* (Schmid, 2013; Hafez and Mullins, 2015), *psychology* (Moghaddam, 2005; Van der Veen, 2016), *computing* (Agarwal and Sureka 2015a), *policing* (Silber et al., 2007), and *governance* (European Parliament, 2019). These efforts however, have mostly evolved in silos, and most of the existing works towards the design and development of AI technology to counter online radicalisation are neither based on, nor do they take advantage of, the existing theories and studies of radicalisation coming from social sciences, psychology or policing.

Models from social science, psychology and policing have investigated the **factors** that drive people to get radicalised (Moghaddam, 2005) (e.g., failed integration, poverty, discrimination),

their different **roots** (Schmid, 2013; Borum, 2016) (micro-level, or individual level, mesolevel, or group/community level, and macro-level, or global level, the influence of government and society at home and abroad), and how the radicalisation process happens and evolves, i.e., what are its different **stages** (Silber et al., 2007) (e.g., pre-radicalisation, self-identification, indoctrination, Jihadisation) . However, very few works in the literature (Lara-Cabrera et al., 2017; Fernandez et al., 2018) have use the learning from these models to create more effective radicalisation analysis and detection methods. AI technology development needs to leverage closer the knowledge of theoretical models of radicalisation to design more effective technological solutions to target online radicalisation.

Adaptation of extremist groups

While multiple efforts are being made to design and develop effective AI solutions that automatically identify and block radical accounts, or that stop the viral spreading or radical content, extremist groups are adapting their behaviour to avoid being detected, or to resurface once they have been blocked (Conway et al., 2017). We list here some of the adaptation techniques used by these groups to maintain their online presence (Bodo, 2018).

- **Content adaptation**: In order to avoid being flagged by AI technology, extremist organisations adapt their content, either by replacing/modifying terms, or by distorting the audio and pixilation of images and videos (Stalman, 2019).
- **User-account adaptation:** Some extremist groups use proxies, such as media organizations or local charities, to post content on the platforms for them to avoid being detected (Frenkel and Hubbard, 2019). In the cases where the accounts are blocked, extremist groups manage to keep reemerging within the same platform under different names, using a variety of strategies to be found by their followers (Conway et al., 2017).
- **Platform adaptation:** In some cases, extremist groups also change platforms. An example is the Islamic State (IS), which in recent years moved many of its communication to Telegram (al-Lami, 2018; Clifford and Powell, 2019) due to the disruption they faced on more visible platforms such as Twitter (Conway et al., 2017), and more recently they are exploring the use of the decentralised Web via platforms like RocketChat and ZeroNet (King, 2019).
- **Technology adaptation:** Extremist organisations make use of the latest technological developments in order to increase the spread of their message. A key example of this is the use of life-stream videos. For example, during the recent attack at Christchurch, New Zealand, the video of the shooting was spread all around the internet. Despite the efforts of tech companies to contain the virality, many hours after the shooting, various clips of the video were still searchable (Lapowsky, 2019).

Ethics and Conflicts in Legislation

Another key challenge of the design, development and use of AI to counter radicalisation is that **technology needs to comply with legislation that can sometimes be ambiguous or contradictory**, particularly when it comes to the tension between security, privacy and freedom of expression. The European Commission, for example, is proposing legislation to ensure all member states bring in sanctions against those who repeatedly fail to respond to removal orders of radical content, facing penalties up to 4% of their global revenue. The draft regulation was approved by Members of the European Parliament (MEPs) in April 2019 (European Parliament, 2019). Critics, including internet freedom think tanks and big tech firms, claim the legislation threatens the principles of a free and open internet (Porter, 2019).[9] Another example is the regulation that will force WhatsApp, Facebook and other social media platforms to disclose encrypted messages from suspected terrorists under a new treaty between the UK and US (Swinford, 2019), with a similar law is already approved in Australia.[10] Privacy advocates are highly critical and have highlighted the potential negative implications that these regulations will have for future cases on privacy and government surveillance.

The other key issue that emerges from the use of AI to counter online radicalisation is the need for a constant review of ethical guidelines in order to assess the risk of the proposed technology and address them through reflexivity and anticipation (Troullinou and d'Aquin, 2018). Processes and decisions, once undertaken by humans, are now computer-driven, increasingly derived through AI powered by big data. And, while reports from the European Union (EU) (European Commission, 2018) state the need of AI to be based on values of respect for human dignity, freedom, democracy, equality, the rule of law, and respect for human rights[11]; the reality is that the rapid and ethically careless development of AI has led to serious adverse effects that go against these values (Harford, 2014). In the case of the development of AI systems to counter online extremism, the wrong categorisation of a user as 'radical' or 'extremist' may result in an innocent person being subjected to surveillance. It is therefore extremely important to consider potential sources of inaccuracy of automatic AI-powered radicalisation detection approaches, and to have a constant reflection and continuous change in ethical guidelines to reduce the potential negative impact of AI developments.

Opportunities

The challenges and issues reported above open a wide range of opportunities for the improvement of AI solutions to counter online radicalisation. We highlight here six main lines of research that we hope to inspire the design and development of future AI technology: (i) stronger collaboration

[9] https://cdt.org/files/2019/02/Civil-Society-Letter-to-European-Parliament-on-Terrorism-Database.pdf
[10] https://www.bbc.co.uk/news/world-australia-46463029
[11] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:12016M/TXTandfrom=EN

across research disciplines and organisations, (ii) creation of reliable datasets to study online radicalisation, (iii) development of comparative studies, (iv) contextual adaptation of technological solutions (v) better integration of humans and technology and, (vi) ethical vigilance.

Collaboration across research disciplines and organisations

Since 2017 several initiatives, such as the Global Internet Forum,[12] or Tech Against Terrorism[13] have emerged, putting Tech Companies in contact with Governments, Civil Societies, researchers and NGOS in order to have a better understanding of what radicalisation is, and how to stop the online phenomenon. These initiatives are helping to create consensus, and to define more clearly what constitutes radical and extremist internet activity, since tech companies should not be the 'deciders' of content moderation (Saltman, 2019). It is also necessary to include different points of view on the discussion table, to ensure the right balance between security, privacy, freedom of expression and content moderation.

While initiatives have emerged to ensure a wide range of organisations are collaborating towards the development of AI solutions to counter radicalisation, synergies across different research fields (psychology, social science, policing, computer science) are yet to become a reality (Scrivens and Davies, 2018). AI design and development can strongly benefit from leveraging closer the knowledge of theoretical models of radicalisation, and the empirical evidence gathered through policing research, to design more effective technological solutions to target online radicalisation.

Creation of reliable datasets to study radicalisation

As we observed in the previous section, the majority of ground truth datasets used to study online radicalisation lack of solid verification. We continue to observe false positives, incompleteness and biases in those datasets. Many datasets used in radicalization studies are no longer available and recollecting that data is no longer possible. Obtaining and annotating data to create reliable gold standard datasets (as well as sharing them for reproducibility purposes) are key future steps for research on online radicalisation.

Comparative Studies

As previously reported, different AI approaches have been developed to counter online radicalisation. However, while some of these approaches target the same objective (e.g., identify radical content / identify radical accounts), they have not been compared against one another. Replication studies are therefore needed to assess existing approaches and techniques, to

[12] https://www.gifct.org/leadership/

[13] https://www.techagainstterrorism.org/

understand their strengths and limitations and to determine which ones should be applied and under which conditions.

Contextual Adaptation of Technological Solutions

As mentioned in our first reported challenge radicalisation needs to be understood in context (time, geographic location, culture, etc.). The same piece of content may be deemed as radical within a particular region of the world, and as non radical in a different region. Similarly, as shown by (Fernandez and Alani, 2018), contextual divergences also emerge within the use of radicalisation terms, and understanding such nuances can help to enhance existing radicalisation detection approaches. It is therefore important to develop robust technological solutions, able to adapt to the different contexts in which they may need to operate.

Better integration of humans and technology

Radicalisation is a human-driven problem, and to develop effective AI solutions to counter this problem it is important to introduce humans in the loop. Human feedback and expertise can be applied at various levels including:

- Co-creation with users: technology development could benefit from the use of co-creation to ensure that different points of view and perspectives are gathered and that this complex problem is targeted simultaneously from different angles.
- Technology and humans deciding together: Expertise may be needed to review complex software decisions. Developing technology that facilitates that human expertise is integrated in the decision-making process could help mitigating the impact of erroneous or controversial outputs.
- Human feedback for technology adaptation: The development of technology that gathers and integrates human feedback can help ensuring that algorithms are retrained**,** capturing evolving behaviours, themes, and novel radicalisation strategies.

Ethical Vigilance

As we previously discussed, there is a strong tension between ensuring security, privacy, and freedom of expression, when targeting online radicalisation. Ethical methodologies are therefore needed to track the human and societal effects of AI technologies during the design, development, and post-production processes. Particularly, the development of postmarket ethical monitoring methods will be needed to further refine, confirm or deny, the safety of a particular technology after it is used to counter online extremism, helping to identify potential unforeseen negative effects.

Conclusions

In this book chapter we have provided an overview of the current AI technological advancements towards addressing the problem of online extremism, identified some of the limitations of existing solutions, and highlighted some opportunities for future research. We hope the provided critical reflections will stimulate discussions on the future design and development of AI technology to target the problem of online extremism.

References

(Agarwal and Sureka, 2015a) Swati Agarwal and Ashish Sureka. 2015a. Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. arXiv preprint arXiv:1511.06858 (2015).

(Agarwal and Sureka, 2015b) Swati Agarwal and Ashish Sureka. 2015b. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In International Conference on Distributed Computing and Internet Technology. Springer, 431–442.

(al-Lami, 2018) Mina al-Lami. 2018. Jihadist media's cat and mouse game. BBC Monitoring. https://monitoring.bbc.co.uk/inside-bbcm/7

(Ashcroft et al., 2015) Michael Ashcroft, Ali Fisher, Lisa Kaati, Enghin Omer, and Nico Prucha. 2015 . Detecting jihadist messages on twitter. In Intelligence and Security Informatics Conference (EISIC), 2015 European. IEEE, 161–164.

(Badawy and Ferrara, 2018) Badawy, A. and Ferrara, E., 2018. The rise of jihadist propaganda on social networks. Journal of Computational Social Science, 1(2), pp.453-470.

(Berger and Strathearn, 2013) JM Berger and Bill Strathearn. 2013 . Who Matters Online: Measuring influence, evaluating content and countering violent extremism in online social networks. International Centre for the Study of Radicalisation and Political Violence (2013).

(Berger and Morgan, 2015) Jonathon M Berger and Jonathon Morgan. 2015. The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter. The Brookings Project on US Relations with the Islamic World 3, 20 (2015), 4–1.

(Bermingham et al., 2009) Adam Bermingham, Maura Conway, Lisa McInerney, Neil O'Hare, and Alan F Smeaton. 2009. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In Int. Conf. Advances in Social Network Analysis and Mining (ASONAM'09).

(Bodo, 2018) Lorand Bodo. Now you see it, now you don't? Moving beyond account and content removal in digital counter-extremism operations. https://www.voxpol.eu/now-you-see-it-now-you-dont-moving-beyond-account-content-removal-in-digital-counter-extremism-operations/

(Borum, 2016) Randy Borum. 2016 . The Etiology of Radicalization. The Handbook of the Criminology of Terrorism (2016), 17.

(Carter et al., 2014) Joseph A Carter, Shiraz Maher, and Peter R Neumann. 2014 . # Greenbirds: Measuring Importance and Influence in Syrian Foreign Fighter Networks. (2014).

(Chatfield et al., 2015) Akemi Takeoka Chatfield, Christopher G Reddick, and Uuf Brajawidagda. 2015. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. In Proceedings of the 16th Annual International Conference on Digital Government Research. ACM, 239–249.

(Clifford and Powell, 2019) Clifford, Bennett and Powell, Helen. Encrypted Extremism. Inside the English-Speaking Islamic State Ecosystem on Telegram. Program on extremism https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/EncryptedExtremism.pdf

(Conway et al., 2017) Conway, Maura, et al. Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts, Studies in Conflict and Terrorism, 42:1-2, 141-160, DOI: 10.1080/1057610X.2018.1513984

(Conway, 2018) Conway, Maura. "Violent Extremism and Terrorism online in 2018. The Year in Review." (2018).

(Correa and Sureka, 2013) Denzil Correa and Ashish Sureka. 2013 . Solutions to detect and analyze online radicalization: a survey. arXiv preprint arXiv:1301.4916 (2013).

(Edwards and Gribbon 2013) Charlie Edwards and Luke Gribbon. 2013 . Pathways to violent extremism in the digital era. The RUSI Journal 158, 5 (2013), 40–47.

(European Commision, 2002) European Commission. Council Framework Decision 2002/475/JHA of 13 June 2002 on Combating Terrorism.

(European Commission, 2018) European Commission. Communication Artificial Intelligence for Europe (2018). https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe

(European Parliament, 2019) European Parliament legislative resolution of 17 April 2019 on the proposal for a regulation of the European Parliament and of the Council on preventing the dissemination of terrorist content online (COM(2018)0640 – C8-0405/2018 – 2018/0331(COD)) https://www.europarl.europa.eu/doceo/document/TA-8-2019-0421_EN.html

(Expert Group, 2008) Radicalisation Processes Leading to Acts of Terrorism. A Concise Report prepared by the European Commission's Expert Group on Violent Radicalisation. Submitted to the European Commission on 15 May (2008).

(Farrel et al., 2019) Farrell, Tracie; Fernandez, Miriam; Novotny, Jakub and Alani, Harith (2019). Exploring Misogyny across the Manosphere in Reddit. In: WebSci '19 Proceedings of the 10th ACM Conference on Web Science, pp. 87–96.

(Farwell, 2014) Farwell, James P. "The media strategy of ISIS." Survival 56.6 (2014): 49-55.

(Fathali and Moghaddam, 2005) Fathali M Moghaddam. 2005. The staircase to terrorism: A psychological exploration. American Psychologist 60, 2 (2005), 161

(Fernandez et al., 2018) Fernandez, Miriam, Moizzah Asif, and Harith Alani. "Understanding the roots of radicalisation on twitter." Proceedings of the 10th ACM Conference on Web Science. ACM, 2018.

(Fernandez and Alani 2018) Fernandez, Miriam, and Harith Alani. "Contextual semantics for radicalisation detection on Twitter." (2018). In: Semantic Web for Social Good Workshop (SW4SG) at International Semantic Web Conference 2018, 9 Oct 2018, CEUR.

(Fernandez et al., 2019) Fernandez, Miriam; Gonzalez-Pardo, Antonio and Alani, Harith (2019). Radicalisation Influence in Social Media. Journal of Web Science. Vol.6

(Ferrara et al., 2016) Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. 2016 . Predicting online extremism, content adopters, and interaction reciprocity. In International conference on social informatics. Springer, 22–39.

(Ferrara, 2017) Emilio Ferrara. 2017 . Contagion dynamics of extremist propaganda in social networks. Information Sciences 418 (2017), 1–12.

(Frenkel and Hubbard, 2019) Sheera Frenkel and Ben Hubbard. 2019. After Social Media Bans, Militant Groups Found Ways to Remain. The New York Times. https://www.nytimes.com/2019/04/19/technology/terrorist-groups-social-media.html

(Gartenstein-Ross and Barr) Daveed Gartenstein-Ross and Nathaniel Barr. 2016. The Myth of Lone-Wolf Terrorism. The Attacks in Europe and Digital Extremism. Foreing Affairs. https://www.foreignaffairs.com/articles/western-europe/2016-07-26/myth-lone-wolf-terrorism

(GDPR, 2019) General Data Protection Regulation. Last access October 2019. https://gdpr-info.eu

(Harford, 2014) Harford, Tim. "Big data: A big mistake?." Significance 11.5 (2014): 14-19.

(Hafez and Mullins, 2015) Mohammed Hafez and Creighton Mullins. 2015 . The radicalization puzzle: A theoretical synthesis of empirical approaches to homegrown extremism. Studies in Conflict and Terrorism (2015).

(Housen-Couriel et al., 2019) Deborah Housen-Couriel, Boaz Ganor, Uri Ben Yaakov, Stevie Weinberg and Dafne Beri. The International Cyber Terrorism Regulation Project. https://rusi.org/publication/other-publications/international-cyber-terrorism-regulation-project

(Kaggle, 2019) Kaggle datasets to study radicalisation. Last accessed October 2019: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter, https://www.kaggle.com/activegalaxy/isis-related-tweets

(King, 2019) Peter King. Analysis: Islamic State's experiments with the decentralised web. BBC Monitoring. https://monitoring.bbc.co.uk/product/c200paga

(King and Taylor 2011) Michael King and Donald M Taylor. 2011 . The radicalization of homegrown jihadists: A review of theoretical models and social psychological evidence. Terrorism and Political Violence 23, 4 (2011), 602–622.

(Klausen, 2015) Jytte Klausen. 2015 . Tweeting the Jihad: Social media networks of Western foreign fighters in Syria and Iraq. Studies in Conflict and Terrorism 38, 1 (2015).

(Kruglanski, 2014) Arie W Kruglanski, Michele J Gelfand, Jocelyn J Bélanger, Anna Sheveland, Malkanthi Hetiarachchi, and Rohan Gunaratna. 2014 . The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. Political Psychology 35, S1 (2014), 69–93.

(Lara-Cabrera et al., 2017) Raúl Lara-Cabrera, Antonio Gonzalez-Pardo, and David Camacho. 2017 . Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter. Future Generation Computer Systems (2017).

(Lapowsky, 2019) Issie Lapowsky. 2019. Why Tech Didn't Stop the New Zealand Attack From Going Viral. WIRED. https://www.wired.com/story/new-zealand-shooting-video-social-media/

(Lygre et al., 2011) Ragnhild B Lygre, Jarle Eid, Gerry Larsson, and Magnus Ranstorp. 2011 . Terrorism as a process: A critical review of Moghaddam's "Staircase to Terrorism". Scandinavian journal of psychology 52, 6 (2011), 609–616.

(Magdy et al., 2016) Magdy, Walid; Darwish, Kareem; Weber, Ingmar. #FailedRevolutions: Using Twitter to study the antecedents of ISIS support. First Monday, (S.l.), jan. 2016. ISSN 13960466.

(Mccauley and Moskalenko, 2008) Clark McCauley and Sophia Moskalenko. 2008 . Mechanisms of political radicalization: Pathways toward terrorism. Terrorism and political violence 20, 3 (2008).

(Meserole and Byman, 2019) Chris Meserole and Daniel Byman. 2019. Terrorist Definitions and Designations Lists. What Technology Companies Need to Know. Global Research Network on Terrorism and Technology: Paper No. 7

(Moghaddam, 2005) Fathali M Moghaddam. 2005. The staircase to terrorism: A psychological exploration. American Psychologist 60, 2 (2005), 161.

(Moskalenko and McCauley, 2009) Sophia Moskalenko and Clark McCauley (2009) Measuring Political Mobilization: The Distinction Between Activism and Radicalism, Terrorism and Political Violence, 21:2, 239-260, DOI: 10.1080/09546550902765508

(O'Callaghan et al., 2014) Derek O'Callaghan, Nico Prucha, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. 2014. Online social media in the Syria conflict: Encompassing the extremes and the in-betweens. In Int. Conf. Advances in Social Networks Analysis and Mining (ASONAM). Beijing, China.

(Olteanu et al., 2019) Olteanu, Alexandra, et al. "Social data: Biases, methodological pitfalls, and ethical boundaries." Frontiers in Big Data 2 (2019): 13.

(Olteanu et al., 2017) Olteanu, Alexandra, Kartik Talamadupula, and Kush R. Varshney. "The limits of abstract evaluation metrics: The case of hate speech detection." Proceedings of the 2017 ACM on Web Science Conference. ACM, 2017.

(Patton et al., 2019) Patton, Desmond U., et al. "Annotating Twitter Data from Vulnerable Populations: Evaluating Disagreement Between Domain Experts and Graduate Student Annotators."

(Parekh et al., 2018) Parekh, Deven, et al. "Studying Jihadists on Social Media: A Critique of Data Collection Methodologies." Perspectives on Terrorism 12.3 (2018): 5-23.

(Porter, 2019) Jon Porter. 2019. Upload filters and one-hour takedows: the EU's latest fight against terrorism online, explained. The Verge. https://www.theverge.com/2019/3/21/18274201/european-terrorist-content-regulation-extremist-terreg-upload-filter-one-hour-takedown-eu

(Rowe and Saif, 2016) Matthew Rowe and Hassan Saif. 2016. Mining Pro-ISIS Radicalisation Signals from Social Media Users. In Int. Conf. Weblogs and Social Media (ICWSM). Cologne, Germany.

(Sageman, 2004) Marc Sageman, Understanding Terror Networks (Philadelphia: University of Pennsylvania Press, 2004), p. 115

(Saif, 2017) Hassan Saif, Thomas Dickinson, Leon Kastler, Miriam Fernandez, and Harith Alani. 2017. A semantic graph-based approach for radicalisation detection on social media. In European Semantic Web Conference. Springer, 571–587.

(Scrivens and Davies, 2018) Ryan Scrivens and Garth Davies. 2018. Identifying radical content online. https://policyoptions.irpp.org/magazines/january-2018/identifying-radical-content-online/

(Schmid, 2013) Alex P Schmid. 2013. Radicalisation, de-radicalisation, counter-radicalisation: A conceptual discussion and literature review. ICCT Research Paper 97 (2013), 22.

(Silber et al., 2007) Mitchell D Silber, Arvin Bhatt, and Senior Intelligence Analysts. 2007. Radicalization in the West: The homegrown threat. Police Department New York.

(Saltman, 2019) Saltman E. 2019: Global Research Network on Terrorism and Technology. https://www.youtube.com/watch?v=82L3ziU4LkM

(Swinford, 2019) Steven Swinford. 2019. Police can access suspects' Facebook and WhatsApp messages in deal with US. The Times. https://www.thetimes.co.uk/edition/news/police-can-access-suspects-facebook-and-whatsapp-messages-in-deal-with-us-q7lrfmchz

(Twitter data policies, 2019) Twitter Data policies and privacy regulations, 2019. https://cdn.cms-twdigitalassets.com/content/dam/legal-twitter/site-assets/privacy-policy-new/Privacy-Policy-Terms-of-Service_EN.pdf, https://twitter.com/en/privacy, https://developer.twitter.com/en/developer-terms/policy

(Tobaili et al., 2019) Tobaili, Taha; Fernandez, Miriam; Alani, Harith; Sharafeddine, Sanaa; Hajj, Hazem and Glavas, Goran (2019). SenZi: A Sentiment Analysis Lexicon for the Latinised Arabic (Arabizi). In: International Conference Recent Advances In Natural Language Processing (RANLP 2019) . pp. 1204–1212.

(Troullinou and d'Aquin, 2018) Troullinou, Pinelopi and d'Aquin Mathieu- Using Futuristic Scenarios for an Interdisciplinary Discussion on the Feasibility and Implications of Technology. Black Mirror and Critical Media Theory, 2018. Rowman and Littlefield

(Van der Veen, 2016) Jaap van der Veen. 2016. Predicting susceptibility to radicalization: An empirical exploration of psychological needs and perceptions of deprivation, injustice, and group threat. (2016).

(Vergani and Bliuc, 2015) Matteo Vergani and Ana-Maria Bliuc. 2015. The evolution of the ISIS'language: a quantitative analysis of the language of the first year of Dabiq magazine. Sicurezza, Terrorismo e Società= Security, Terrorism and Society 2, 2 (2015), 7–20.

(Von Behr et al., 2013) Ines von Behr, Anaïs Reding, Charlie Edwards, Luke Gribbon. 2013. Radicalisation in the digital era: The use of the Internet in 15 cases of terrorism and extremism. (2013). https://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR453/RAND_RR453.pdf